# Conifer: Clonal Tree Inference for Tumor Heterogeneity With Single-cell and Bulk Sequencing Data

**Leila Baghaarabani**
Institute of Biochemistry and Biophysics, University of Tehran
**Sama Goliaei**
New Sciences and Technologies, University of Tehran
**Mohammad-Hadi Foroughmand-Araabi**
Department of Mathematical Sciences, Sharif University of Technology
**Seyed Peyman Shariatpanahi**
Institute of Biochemistry and Biophysics, University of Tehran
**Bahram Goliaei** ( goliaei@ut.ac.ir )
Institute of Biochemistry and Biophysics, University of Tehran

**Research Article**

**Keywords:** Heterogeneity of tumor, Clonal tree, Bulk sequencing, single-cell sequencing, Heterogeneity of tumor, Bayesian nonparametric model

**Posted Date:** March 1st, 2021

**DOI:** https://doi.org/10.21203/rs.3.rs-263502/v1

# Conifer: Clonal Tree Inference for Tumor Heterogeneity with Single-Cell and Bulk Sequencing Data

**Leila Baghaarabani[1], Sama Goliaei[2], Mohammad-Hadi Foroughmand-Araabi[3], Seyed Peyman Shariatpanahi[1], Bahram Goliaei[1*]**

[1]Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran.

[2]Faculty of New Sciences and Technologies, University of Tehran, Tehran, Iran.

[3]Department of Mathematical Sciences, Sharif University of Technology, Tehran, Iran.

***Corresponding Author**

Bahram Goliaei, Ph.D.,

Institute of Biochemistry and Biophysics

University of Tehran, Tehran, Iran.

E-mail: goliaei@ut.ac.ir

## Abstract

**Background:** An important and effective step in cancer treatment is understanding the clonal evolution of cancer tumors. Clones are cell populations with different genotypes, resulting from the differences in the somatic mutations that occur and accumulate during cancer development. An appropriate approach for better understanding a tumor population is determining the variant allele frequency with which the mutation occurs in the entire population. Bulk sequencing data can be used to provide that information, but the frequencies are not informative enough in identifying different clones and their evolutionary relationships. On the other hand, single-cell sequencing data provides valuable information about branching events in the evolution of a cancerous tumor. However, in the single-cell sequencing data, the total population of sequenced cells is naturally much smaller than bulk sequencing so it is not precise enough for calculating cell prevalence.

**Result:** In this study, a new method called Conifer (ClONal tree Inference For hEterogeneity of tumoR) is proposed which combines aggregated variant allele frequency from bulk sequencing data with branch evolution information from single-cell sequencing data, in order to better understand clones and their evolutionary relationships. It is proven that the accuracy of clone identification is increased by using Conifer compared to other existing methods in both real and simulated data. Also, it is shown that the approach of Conifer in using single-cell sequencing data together with bulk sequencing data has reduced the possibility of cloning mutations with similar frequency but belonging to different clones.

**Conclusions:** In this study, we provided an accurate and robust method to identify clones of tumor heterogeneity and their evolutionary history by combining single-cell and bulk sequencing data.

***Keywords:*** *Heterogeneity of tumor, Clonal tree, Bulk sequencing, single-cell sequencing, Heterogeneity of tumor, Bayesian nonparametric model.*

## Background

Genetic mutation is a major cause of abnormal cell growth and cancer. Although cancer cells are usually derived from one mutated cell initially [1] and therefore have shared mutated genes, new mutations may happen further in cancer development. In other words, the cancer cells in a tumor are not homogeneous and tumor genomic heterogeneity is already shown in many studies [1-3]. A tumor consists of different clones, where each clone is a set of cells sharing a common genotype inherited from a common ancestor [4]. For effective treatment of cancer, it is critical to diagnose clones of a tumor, to determine the development stage of cancer cells, and to identify early single-nucleotide variants (SNVs) that have led to rapid cell growth and multiplications. Consequently, identifying tumor heterogeneity and its phylogenetic inference is an essential step in effective cancer treatment [5-7].

Sequencing of bulk data that focuses on DNA of mixture of thousands or millions of cancerous and/or normal cells is widely used for providing a mixed signal of variant allele frequencies (VAFs) for each somatic mutation. In order to discover the evolutionary history, bulk sequencing data needs deconvolution analysis [8], which often includes two subsequent deduction steps. At the first step, SNV clusters occurring together are deduced by deconvolving the mixed signal of the bulk sample [9]. Afterward, the evolutionary relationship between clusters is deduced by using SNV cluster frequencies [10]. However, in some methods such as PhyloWGS [11], these two inference steps are carried out jointly to avoid SNV clusters that are phylogenetically incompatible.

In most tumor heterogeneity analyses based on bulk sequencing data such as PyClone [9], PhyloSub [12], Clomial [13], and AncesTree [14], it is generally supposed that SNVs with similar VAFs belong to the same clone, which means those SNVs have equal cellular prevalences.

It is shown that in some cases only relying on frequencies observed in a bulk sample may not be enough to infer the evolutionary history, and taking multiple samples is required [8]. In addition, the assumption that SNVs with similar frequency belong to the same clone may be violated, since a tumor may be composed of clones with similar frequencies but different genotypes. Moreover, even though low-frequency SNVs are common in tumor diversity and play a decisive role in treatment, they may be ignored in the process of obtaining the prevalence of SNVs in bulk sequencing [15].

To achieve higher resolution for inferring evolutionary history, single-cell sequencing was introduced, which allows direct acquisition of cell genotypes without the need to deconvolution of mixed signals [16-20], which resulted in reducing the possibility of ignoring low-frequency SNVs. In addition, single-cell information about co-occurred SNVs can be used to differentiate between clusters of SNVs with the same prevalences [21].

Single-cell sequencing is well used in methods such as SCITE [22], OncoNEM [23], and SiFit [24] to infer mutational phylogenetic trees, however, clonal frequencies are not reported in them. Furthermore, in SiCloneFit [25], a nonparametric Bayesian mixture model based on a Chinese Restaurant Process (CRP) is introduced to resolve the clonal genotypes and their evolution.

The single-cell sequencing approach is costly and error-prone. False-positive errors occur due to the DNA amplification error, and false-negative errors occur due to the loss of one or both alleles (dropout). Furthermore, another type of noise may occur in data as a result of accidental sequencing of two or more cells. Moreover, in single-cell sequencing, the number of processed cells is significantly less than bulk sequencing, and also it is naturally inaccurate in calculating cell prevalences.

Considering the advantages and disadvantages of bulk and single-cell sequencing data, the idea of utilizing both data types is incorporated in a number of studies, in order to reduce inaccuracies in each approach and consequently to achieve a better understanding of subclones in cancer tumors.

ddClone [26] analyzes intra-tumor heterogeneity using single-cell and bulk sequencing data and proposes a probabilistic model based on the nonparametric Bayesian method to deduce tumor clones. The prior of the Bayesian method is obtained from single-cell data, and the likelihood is obtained from bulk sequencing data. However, ddClone does not infer tumor phylogeny and is not sufficient for understanding cancer tumor evolution.

B-SCITE [21] is the first computational approach that infers tumor phylogeny from combined single-cell and bulk sequencing data. This probabilistic method searches for tumor phylogenetic trees to maximize the joint likelihood of the two data types. In this method, tree search is carried out with a customized Markov chain Monte Carlo (MCMC) search over the space of labeled trees [21].

In this paper, we propose a new method Conifer, which incorporates both single-cell and bulk sequencing data to infer evolutionary histories of the tumors. Conifer provides clonal genotypes and phylogeny of clones, as well as cell population. We use single-cell sequencing data to resolve the challenge of identifying similar prevalent clones in the tumor and moreover, to resolve ambiguities in the phylogeny inference. On the other hand, we use bulk sequencing data in Conifer to reduce the negative effects of sampling biases and false-negative mutations. Conifer is the first method based on our knowledge that introduces tumor clonal trees using both single-cell and bulk sequencing data.

Conifer is a Bayesian nonparametric model since clones and their evolutionary trees are not predefined, a tree-structure Chinese Restaurant Process (CRP) is used as a prior. To approximate the posterior of the Bayesian model, the particular MCMC algorithm that Conifer performs is a Collapsed Gibbs Sampling in which some of the latent variables are marginalized out to speed up the coverage of the chain. As a result, Conifer introduces a clonal tree in which each node represents the clonal genotypes that have occurred together and are shared between different cells. In nodes closer to the tree root, corresponding clonal genotypes are shared between more cells while by moving in the tree from the root towards its leaves, clones become more specialized to particular cells in those paths.

We evaluated Conifer comprehensively on the various simulated dataset with different numbers of clone, bulk sample, and single-cell sample, etc. and compared it with best methods such as B-SCITE [21] and ddClone [26] in the field, based on criteria like V-measure [27] and adjusted rand score [28] for evaluation of clone identification. Moreover, Conifer introduces the clonal evolutionary trees on simulated data which is significantly similar to the actual trees in different criteria like Common Ancestor Set Distance (CASet), Distinctly Inherited Set (DISC) [29], and Ancestor-Descendant accuracy [21] for various type of datasets. Also, on real cancer data, Conifer provides the evolutionary tree with good agreement with other evidence. Accordingly, Conifer has higher accuracy in clone identification and phylogeny inference than other existing methods in a robust way.

## Results and Discussion

In order to overcome the challenge of identification of clones with similar frequency in bulk sequencing data, and also the challenge of less accuracy of noisy single-cell sequencing data, Conifer combines both data types for inferring clonal phylogeny and its cell prevalence. Furthermore, a clonal tree is introduced by using the Bayesian approach and specifying a generative probabilistic model for hierarchical structure. In this clonal tree, each node is associated with a clone genotype which is a distribution on SNVs.

A single cell mutation profile is generated by choosing a path from the root to a leaf, and repeatedly sampling clones along that path, and sampling SNVs from the selected clones. This generative process uses the probabilistic graphical model shown in Fig. 1. This model is described briefly in this section and more details about hyper-parameters and variables are provided in the Material and method section.

Cell populations, which are collections of cells with almost similar genotypes, are provided by attaching cells to each clonal genotype at different stages of evolution. Conifer method clusters common SNVs in single-cell and bulk sequencing data and introduces the evolutionary tree of clusters. In Fig. 2, it is shown that how two datasets are connected to each other in order to infer the clonal tree.

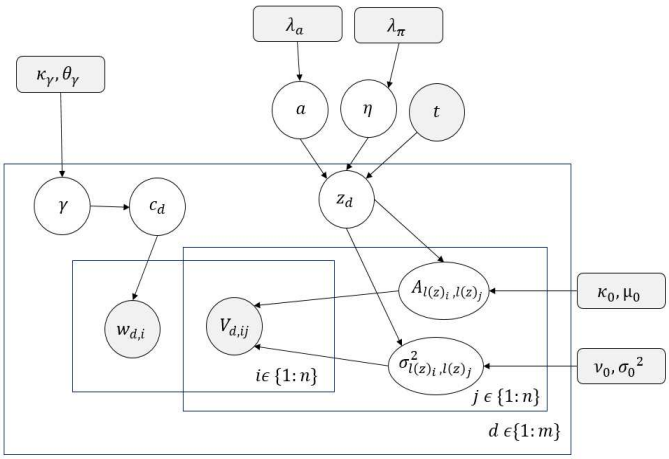



**Fig. 1** Probabilistic graphical model of Conifer. $c_d$ and $z_d$ are path and level assignment generated by tree-structure CRP for mutations of the cell d . $w_d$ is the n-dimensional vector for SNVs which shows whether or not each site is mutant in cell d. $V_d$ is a vector that presents the distance of SNVs which is calculated by VAFs in different bulk samples.

As it is shown in Fig. 2a, single-cell data is represented as a matrix with rows and columns showing SNVs and cells, respectively, and each element indicating the presence or absence of corresponding SNV in a cell. Moreover, bulk sequencing is considered as a matrix in which each element presents VAFs related to SNV in different bulk sequencing samples (Fig. 2b).

In Conifer method, it is assumed that SNVs with similar VAFs in different samples most likely belong to a common cluster, unless there is no single-cell profile that shows two SNVs co-occurred. In fact, all co-occurred patterns in the single-cell profile are considered as the prior knowledge, therefore, with single-cell sequencing, a layout is defined for clustering. In Figure 2c the co-occurred patterns are shown as dashed rectangles. Afterward, the clonal tree is inferred in such a way that its root is a normal genotype (not mutated) and its nodes are clones with the possibility that a set of cell population can be attached to them. The technique of clustering and phylogenetic inference is described in detail in the Material and method section.

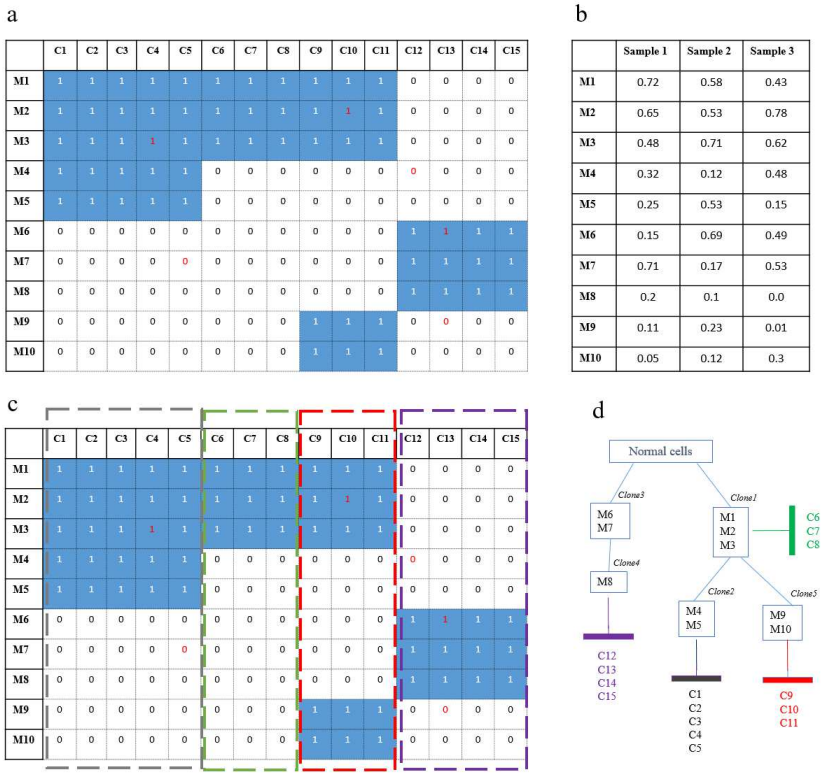



**Fig. 2** Schematic representation of combining single-cell and bulk sequencing data for clonal tree inference in Conifer method. **a)** n × m matrix in which each row and column represents SNVs and cell, respectively. White elements show no mutation and blue ones show mutation has occurrence. 1 and 0 with the red font show false-positive and false negative (drop-out events), respectively. **b)** n × b matrix that its rows are SNVs and its columns are bulk samples and $B_{ij}$ is variant allele frequency in bulk samples. **c)** co-occurred patterns of SNVs in single-cell profiles which are determined by dashed rectangles. **d)** the clonal tree and cell population attachment.

## Performance on simulated data

Since the clonal tree is not known for data of real cancer tumors, in order to evaluate the performance of Conifer method, a complete set of data is simulated and used. To simulate data, the idea of ddClone [26] and B-SCITE [21] studies are used. The simulated data covers various cell counts (50 and 100), number of clones (6, 10, 15, and 20), number of bulk sampling (1,2 and 3), and types of errors in single-cell data. The clonal trees of tumor evolution are produced for each number of genotype clusters, and in each tree, the root node represents a healthy cell population, and SNVs are randomly distributed between other nodes. Details of the model and implementation are given in ddClone [26] and B-SCITE [21].

For evaluating the accuracy of clustering, the V-measure [27] and the adjusted rand score [28] criteria are used which are implemented in scikit-learn Python package 0.19.2. Their corresponding scores are between 0 and 1, which 0 represents random labeling independent of the number of clusters and 1 shows the accurate clustering. Moreover, the inferred clonal tree is evaluated by quantitatively comparing it with the actual tree. The common ancestor of all SNV pairs and also clustering sets (clones) are compared in Conifer and actual trees by assessing the CASet and DISC, respectively [29]. In addition to, the Ancestor-Descendant accuracy [21] is used for evaluation of Ancestor-Descendant relation in Conifer's trees and actual trees. For presenting the accuracy of the methods on different criteria the plots were generated by Ggplot2 [30].

In order to measure method sensitivity respecting errors in single-cell sequencing data, different types of errors are examined that two most important ones are briefly stated as follows:

- Assortment bias: which is a single cell sequencing error that occurs when genotypes of sampled cells do not properly represent the genotypic distribution of tumor cell population. To simulate assortment bias error in single-cell data, new genotype prevalence is obtained by sampling from a Dirichlet distribution with parameter $\lambda$ on the average cell prevalence of bulk sequencing data. Large values of $\lambda$ indicate less assortment bias and equivalently less difference between single-cell and bulk genotype prevalence. In this study, for measuring the sensitivity of Conifer method with respect to assortment bias, four sets of cells with different $\lambda$ ($\lambda$ =1, 5, 10, and 1000) are simulated.

- Doublet: This type of error occurs in single-cell sequencing data when one or more single cells are placed together in a sequencing well, and consequently, their genotypes are mixed with each other, and the signal of a genotype shows a greater mutant locus than each cell trapped in the well. These are considered false-positive errors [26]. For considering this type of error while simulating the single-cell data with probability δ, it is unified with the next simulated cell.

**Clone identification accuracy**: Accuracy of Conifer, ddClone, and B-SCITE methods in SNV clustering is evaluated and compared in Figures 3-6, for 100 clonal trees simulated with 6 and 10 clones, 50 SNVs, 1 to 3 bulk sequencing samples with coverage 10,000, and 50 single-cell genotypes. Simulated single-cell data is generated with the following errors: the false-positive rate of $10^{-5}$, the false-negative rate of 0.2, the missing rate of 0.05, and the doublet rate of 0.2.

According to Fig. 3 and with λ=1 and 10 clusters, Conifer has the mean V-measure of 0.95± 0.05 which shows better performance in clustering than ddClone and B-SCITE with mean V-measure of 0.57± 0.11 and 0.84 ± 0.06, respectively. Additionally, the mean of adjusted rand score for Conifer is 0.86 ± 0.10 which outperforms both ddClone (0.26 ± 0.13) and B-SCITE (0.63 ± 0.14).

It is shown in Fig. 3 that by increasing the value of λ to 5 and 10, Conifer still has more accurate result with mean V-measure of 0.93 ± 0.04 and 0.94 ± 0.05 for λ = 5 and 10, respectively, comparing to the mean of V-measure of ddClone (0.55 ± 0.11 for λ=5, 0.53 ± 0.1 for λ=10) and B-SCITE (0.84 ± 0.06 for λ=5, 0.85 ± 0.07 for λ=10). In addition and according to the adjusted rand score, B-SCITE (0.64 ± 0.11 for λ=5, 0.62 ± 0.09 for λ=10) is more accurate than ddClone (0.24 ± 0.17 for λ=5, 0.22 ± 0.16 for λ=10) and Conifer (0.91 ± 0.3 for λ=5, 0.93 ± 0.1 for λ=10) is the most accurate one.

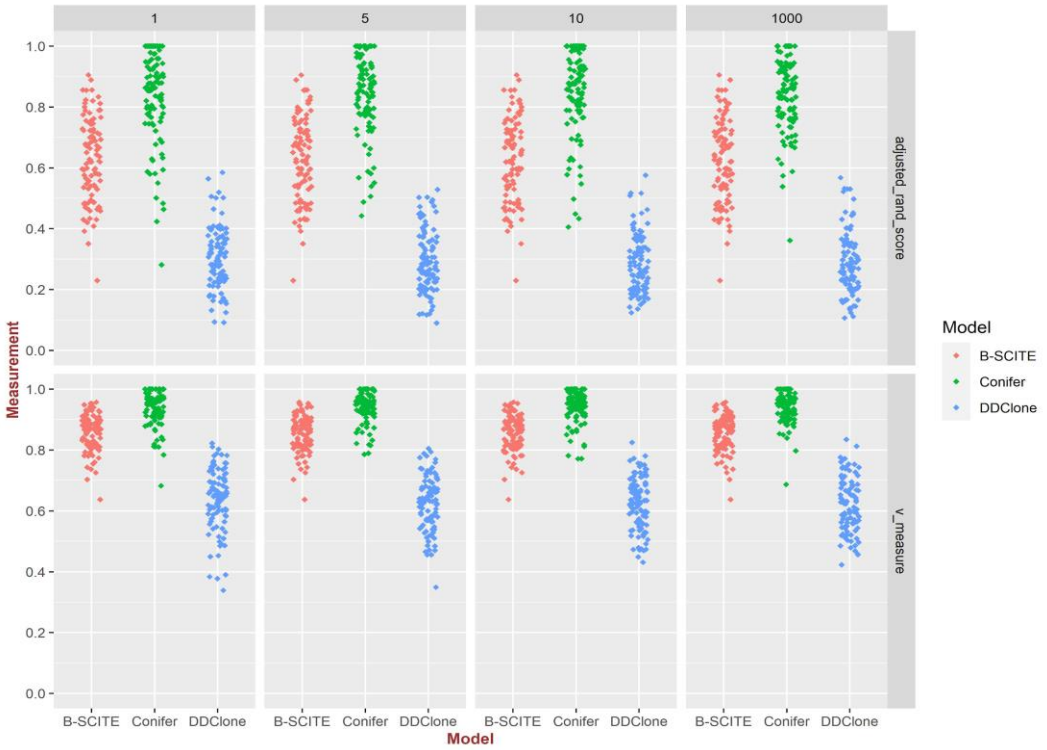



**Fig. 3** Comparison of mutation clustering accuracy in ddClone, B-SCITE, and Conifer methods for 100 clonal trees simulated with 10 clones and 50 mutations. For $\lambda$ = 1,5,10 and 1000. For single-cell data, 50 genotypes are extracted for each clonal tree. The number of bulk sequencing samples is 1 with coverage 10,000. The following errors are added to the single-cell set: the false-positive rate of $10^{-5}$, the false-negative rate of 0.2, missing rate of 0.05, and doublet rate of 0.2.

For $\lambda$=1000 which indicates the least assortment bias used in this study, the resulted mean V-measure and adjusted rand score in Conifer is 0.94 ± 0.06 and 0.91 ± 0.13, respectively. These measures for B-SCITE are 0.85 ± 0.03 and 0.61 ± 0.10 and for ddClone are 0.53 ± 0.05 and 0.23 ± 0.17, which again shows better performance of Conifer method.

To examine the accuracy of the method for different numbers of clusters, it is changed to 6 clusters and the rest of the parameters are kept constant and clustering is repeated for all three methods. As it is shown in Fig. 4, while B-SCITE is more accurate than ddClone, Conifer has considerably outperformed both methods.

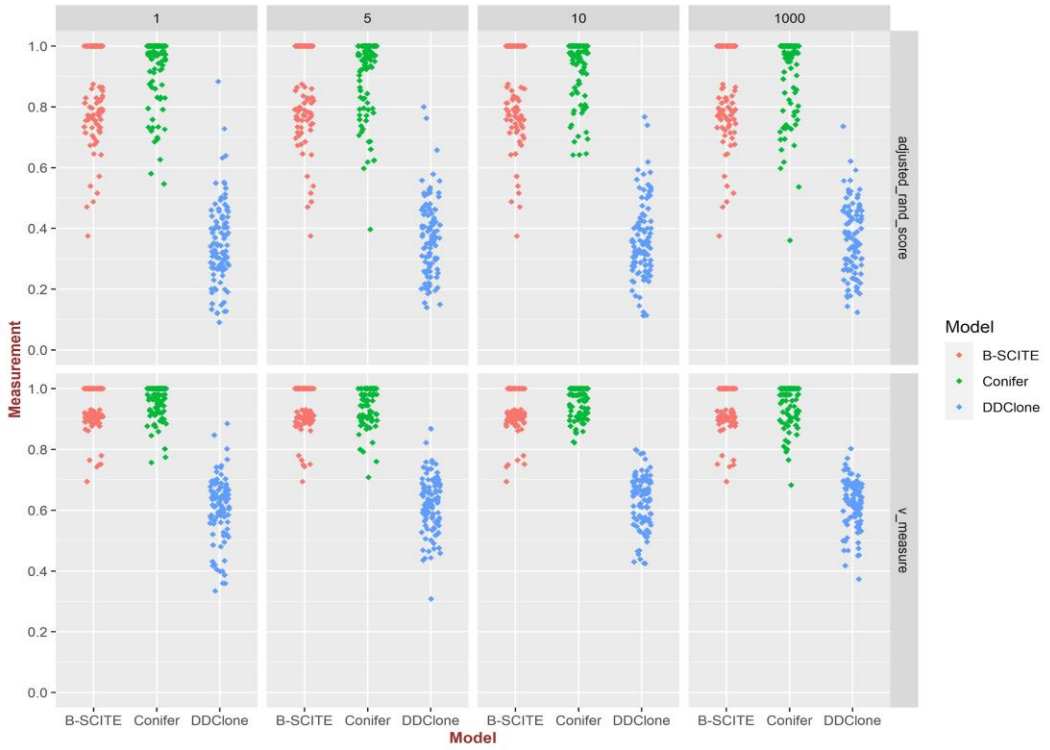



**Fig. 4** Comparison of mutation clustering accuracy in ddClone, B-SCITE, and Conifer methods for 100 clonal trees simulated with 6 clones and 50 mutations. For λ = 1,5,10 and 1000. For single-cell data, 50 genotypes are extracted for each clonal tree. The number of bulk sequencing samples is 1 with coverage 10,000. The following errors are added to the single-cell set: the false-positive rate of $10^{-5}$, the false-negative rate of 0.2, missing rate of 0.05, and doublet rate of 0.2.

**Several bulk samples:** To evaluate the efficiency of Conifer method in several bulk samples, different numbers of bulk sequencing data with 10000 coverage are generated. Since ddClone only considers one bulk sequencing sample, the results are only compared with B-SCITE. As it is shown in Fig. 5, with λ=1 and two bulk samples with coverage 10000, the mean V-measure for Conifer is 0.95 ± 0.06 which is much closer to the perfect clustering comparing to B-SCITE with the mean V-measure of 0.68 ± 0.09. Furthermore, the mean adjusted rand score of B-SCITE is 0.45 ± 0.13 which is less than Conifer's which is 0.86 ± 0.15. According to Fig. 5, Conifer method is still more precise than B-SCITE for other values of λ=5, 10, and 1000.

Regarding the number of bulk samples, the quality of clustering is not change much for both methods by increasing the number of samples and with 3 bulk samples, Conifer remains more accurate than B-SCITE (Fig. 6).

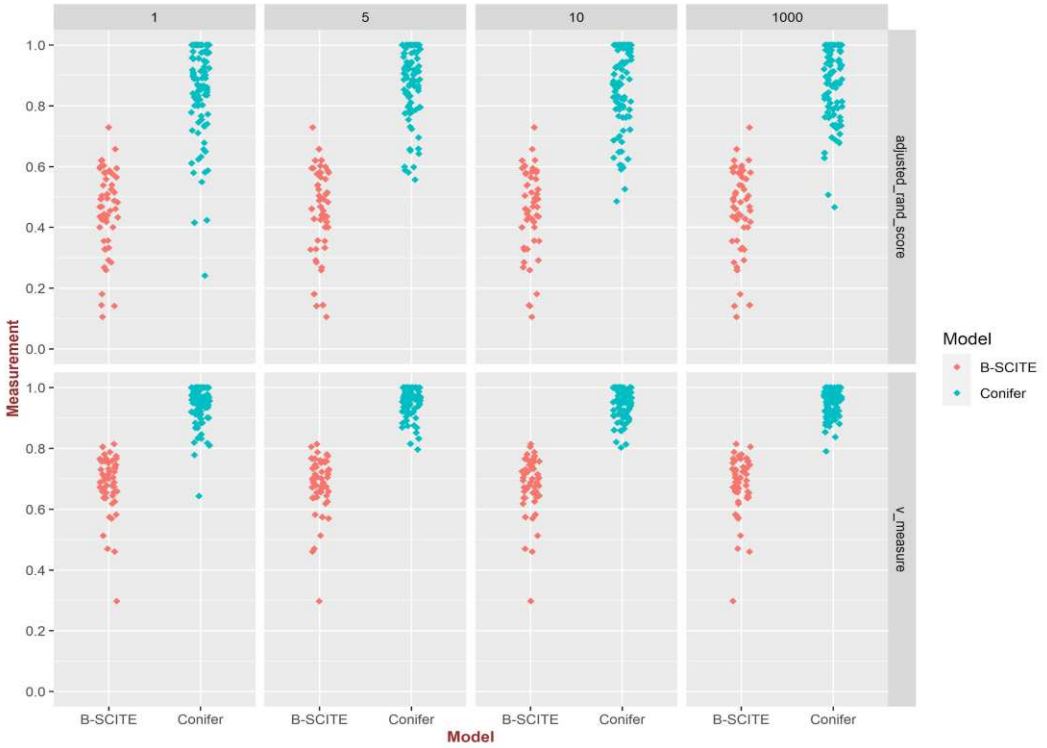



**Fig. 5** Comparison of mutation clustering accuracy in B-SCITE and Conifer methods for 100 clonal trees simulated with 10 clones and 50 mutations. For λ equals 1,5,10 and 1000. For single-cell data, 50 genotypes are extracted for each clonal. The number of bulk sequencing samples is 2 with coverage 10,000. The following errors are added to the single-cell set: the false-positive rate of $10^{-5}$, the false-negative rate of 0.2, missing rate of 0.05, and doublet rate of 0.2.

**Clonal tree accuracy:** In addition to clone identification, since the clonal tree is not provided in ddClone and on the other hand, B-SCITE is mainly designed to infer mutational tree, the accuracy of Conifer clonal tree is not compared to them.

In Fig. 7, the CASet, DISC, and difference in Ancestor-Descendant are shown for 100 clonal trees simulated with 10,15 and 20 numbers of clone, 100 SNVs, 100 genotypes, and 3 bulk sequencing samples with coverage of 10000. Simulated single-cell data is generated with the following errors: the false-positive rate of $10^{-5}$, the false-negative rate of 0.2, missing rate of 0.03, and doublet rate of 0.2. For λ=1, the mean CASet distance which is related to common ancestor for 10, 15 and 20 clones are 0.063 ± 0.079, 0.06 ± 0.075 and 0.056 ± 0.082, respectively. The DISC distance values for 10 ,15 and 20 clones are 0.044 ± 0.052, 0.042 ± 0.039 and 0.043 ± 0.045, respectively which in both cases are smaller than CASet distance. Also the difference in Ancestor-Descendant for 10,

15 and 20 clones are 0.069 ± 0.090, 0.046 ± 0.064 and 0.035 ± 0.053, respectively. It is worth mentioning that increasing the value of λ (to have less assortment bias) has no significant effect on these distances. As a result, the low value of distance criteria from the actual tree indicates the high accuracy of Conifer in tree inference.

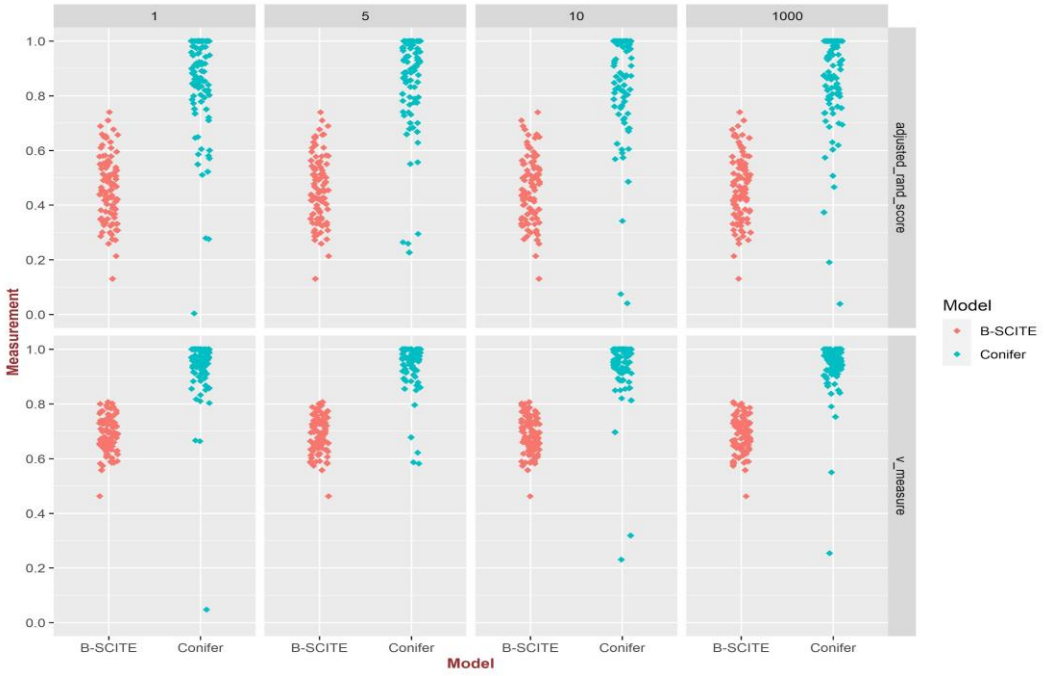



**Fig. 6** Comparison of mutation clustering accuracy in B-SCITE and Conifer methods for 100 clonal trees simulated with 10 clones and 50 mutations. For λ equals to 1,5,10 and 1000. For single-cell data, 50 genotypes are extracted for each clonal tree. The number of bulk sequencing samples is 3 with coverage 10,000. The following errors are added to the single-cell set: the false-positive rate of $10^{-5}$, the false-negative rate of 0.2, missing rate of 0.05, and doublet rate of 0.2.

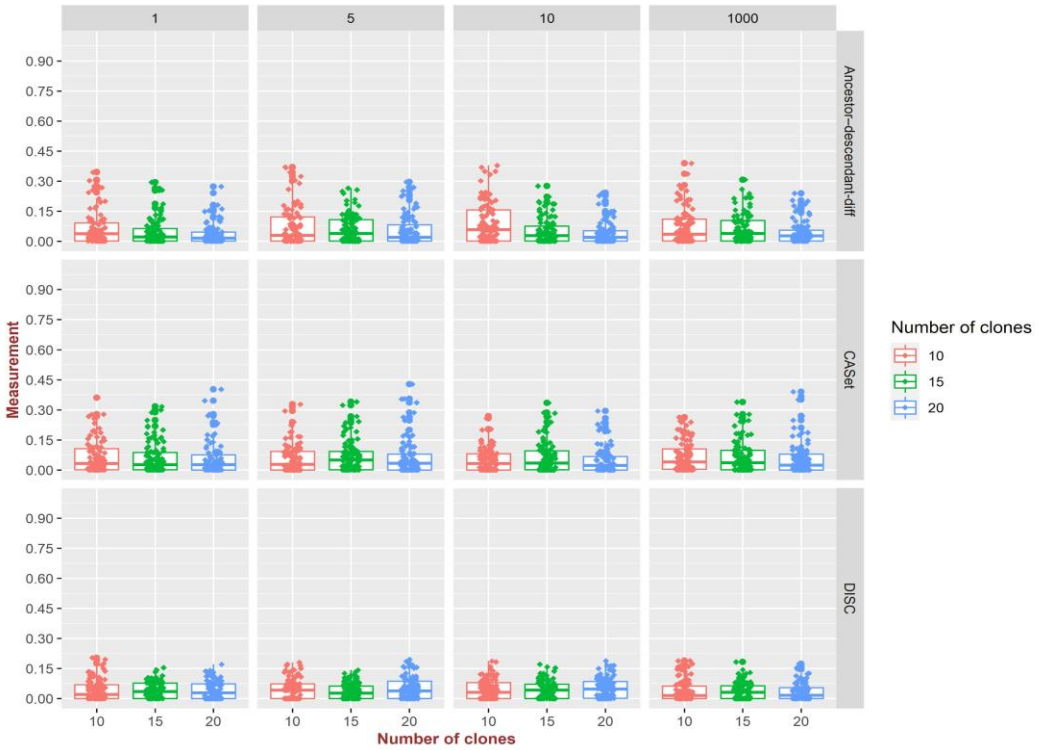



**Fig. 7** The clonal tree distances of Conifer for 100 clonal trees simulated with 10,15 and 20 clones with 100 mutations For $\lambda$ = 1,5,10 and 1000. For single-cell data, 100 genotypes are extracted for each clonal tree. The number of bulk sequencing samples is 3 with coverage 10,000. The following errors are added to the single-cell set: the false-positive rate of $10^{-5}$, the false-negative rate of 0.2, missing rate of 0.03, and doublet rate of 0.2.

**The present of CNV:** Conifer assumes that SNVs are obtained from the copy-number-neutral regions and the VAFs are not affected by copy number alterations. However, Conifer is still robust to alteration of CNV, on account of the fact that single-cell sequencing data is also used for clone identification and tree inference. To examine the accuracy of tree inference in presence of CNV in Fig. 8, it is assumed that 15 percent of SNVs are selected from copy number change regions. For different numbers of Clones and various values $\lambda$ Conifer is quite robust according to this effect and the average of CASet, DISC, and Ancestor-Descendant-diff criteria increased slightly which shows a slight decrease in the accuracy.

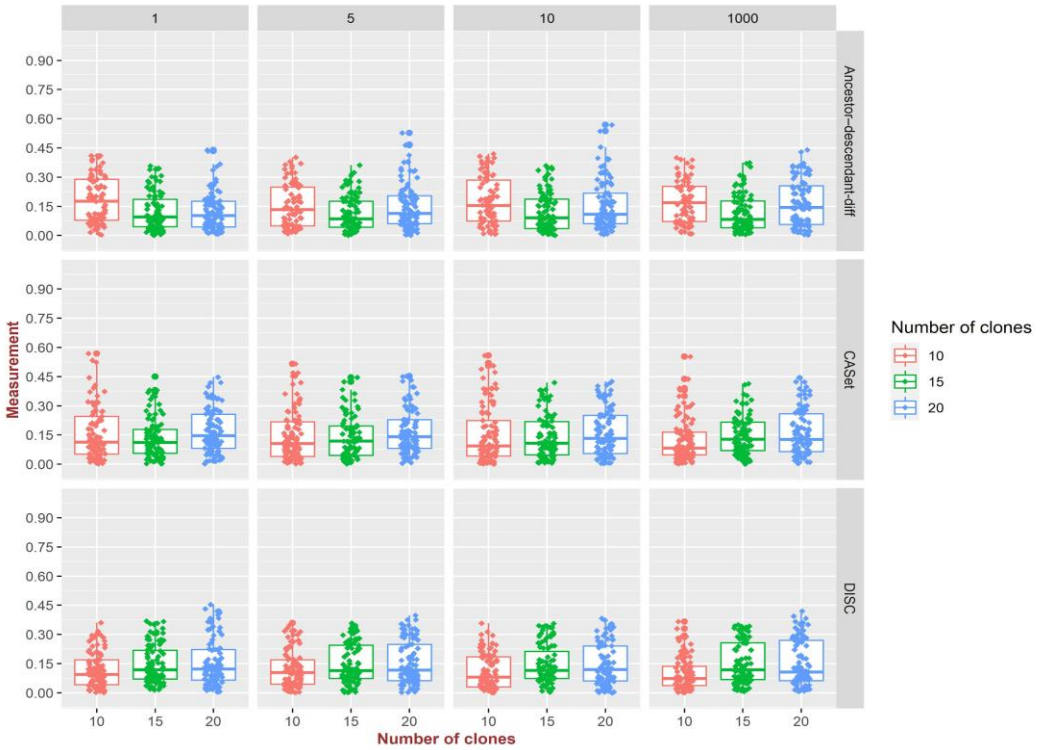



**Fig. 8** The clonal tree distances of Conifer for 100 clonal trees simulated with 10,15 and 20 clones with 100 SNVs for λ = 1,5,10 and 1000. 15 percent of SNVs are selected from copy number change regions. For single-cell data, 100 genotypes are extracted for each clonal tree. The number of bulk sequencing samples is 3 with coverage 10,000. The following errors are added to the single-cell set: the false-positive rate of $10^{-5}$, the false-negative rate of 0.2, missing rate of 0.03, and doublet rate of 0.2.

**Performance on real data:** Conifer performance is further evaluated on real data of a patient (CRC2) with colorectal cancer which is provided in the study of Leung [31]. The noteworthy point in this dataset is the existence of two bulk sequencing data of primary and metastatic tumors together with single-cell sequencing data.

For CRC2 patient, 182 cells are sequenced from the primary colorectal and liver metastatic tumors. The number of SNVs which are reported by the original study is 36. Genotypes reported as binary values indicate the presence or absence of mutation in an SNV locus. In this study, cells with no mutation are eliminated and 25 SNVs and 86 cells are considered for CRC2 patient.

The clonal tree of this dataset inferred by Conifer is shown in Figure 9. Each branch in the tree represents the mutation profile of one or a set of cells, and each clone is a set of mutations that have occurred in a branch and their VAF frequency in different bulk sequencing samples are similar.

Conifer method introduces a tree with 7 nodes (clones) so that the root node is for the non-mutant genotype (C1) and two nodes C2 and C3 are its descendants. C2 is a cluster that contains somatic mutations (SPEN_1, NR3C2, EPHB6, ATR). It is different from primary and metastatic tumor clones and has a separate branch in the clonal tree. This clone and its separated branch are also mentioned in the original study of Leung [31]. C3 is the first evolved clone from healthy cells and has nine mutations, including tp53 which is a tumor suppressor gene. The cell population CP3 is attached to C3 clonal genotype.

In the evolutionary process after C3 clone formation and before tumor metastasis, C4 clone is formed with mutations (CHN1, ATP7B, APC: 1, LRP1B) and CP4 cell population. This clone is introduced as a result of VAFs similarity and mutation occurrence in one single cell profile. In the original study of Leung [31], SCITE method [22] is used to infer the evolutionary tree of mutation, and two distinct branches for metastatic cells are reported based on single-cell sequencing data, assuming that the mutations are not lost during evolution.

Based on both bulk and single-cell sequencing data, Conifer method shows that a group of cells of the primary clone C4 has migrated to the liver, and this migration has occurred only once. Conifer concludes that the migrated cells are subjected to the FUS mutation in the liver creating a clone called C5 and then evolved into two separated branches. The reason that Conifer represents the FUS mutation as a separated clone is that although the FUS mutation should belong to the C6 clone considering neighboring mutations with close frequencies based on single-cell data, as the VAF is not similar to other mutations in that clone (NR4A3, HELZ, TSHZ3), a separated cluster is created. As it is shown in Figure 9, Conifer concludes that in addition to the C6 clone, the C5 clone is also the ancestor of the C7 clone, which can be explained by the false positives that occurred in the profile of eight cells. In fact, it indicates the possibility of co-occurrence of FUS mutation with mutations of C7 (SPEN_2, F8, LAMB4).

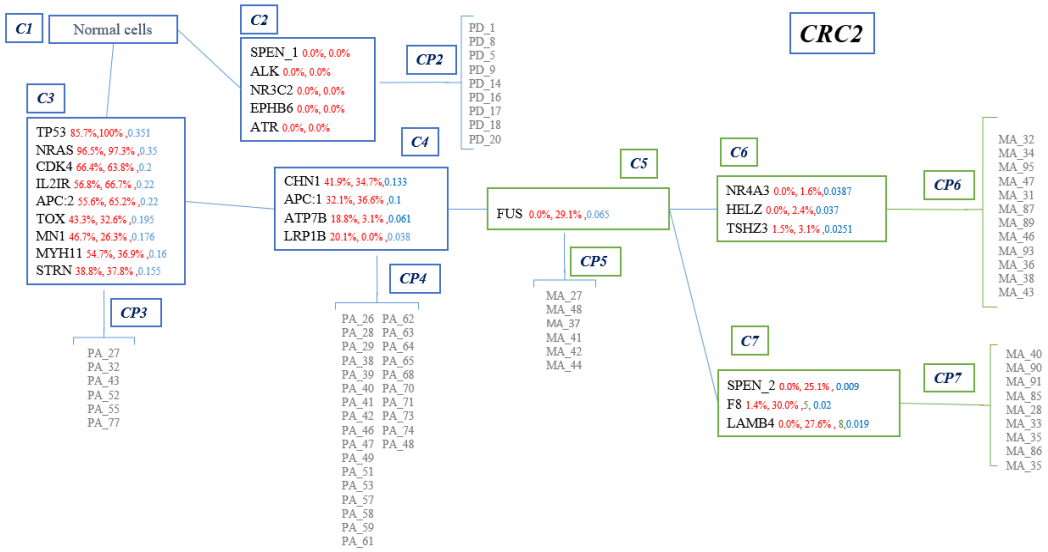



**Fig. 9** Clonal evolution tree for CRC2 patient tumor data. For each SNV, three numbers are reported, from left, the first and second numbers are the VAFs in colorectal tumor bulk sample and metastasis liver bulk sample, respectively, and the third number is the frequency of that SNV in the single-cell sequencing data

Comparison of Conifer inferred tree and the clonal tree introduced in SiCloneFit [25], which is based on only single-cell sequencing data, shows some worth-mentioning differences. In SiCloneFit [25] two "IL2IR" and "APC: 2" mutations are co-occurred in the first clone of the primary tumor. On the contrary and according to the similarity of those two mutations VAFs, Conifer concludes that they belong to the second clone of the primary tumor. Additionally, the clonal tree introduced in SiCloneFit [25] for CRC2 patient represents polyclonal seeding. In other words, it shows the existence of two distinct branches for metastasis. In fact, in SiCloneFit it is concluded that two distinct groups of cells with different mutations have migrated from the primary clone and formed two independent metastatic clones, and the FUS mutation has occurred in both of them independently and during two different evolutionary processes. The Conifer inferred tree in which the creation of FUS mutation occurs only once and before the branching which happens after migration, is more likely to be correct as the VAFs of FUS mutation in the metastatic sample (29.1) is approximately equal to the total mean VAFs of C6 (2.36) and C7 (27.56). A recent study that proposes a method named SCARLET [32] also shows monoclonal seeding by investigating changes in copy number variation of single-cell sequencing data, which validates the Conifer tree.

## Conclusion

In this study, a new reliable and effective method named Conifer is introduced for reconstructing tumor clonal tree by combining single-cell and bulk sequencing data, which can potentially play a crucial role in effective cancer treatment. Conifer provides a generative nonparametric model for clone identification and its evolutionary relationship based on single-cell and bulk sequencing data by considering finite site assumption.

Conifer method has the distinctive feature of simultaneously identifying both clones and phylogenetic tree by combining bulk and single-cell sequencing data. Each tree branch contains mutations of one or more cells, and their common clones are obtained by the VAFs similarity of them in different bulk sequencing samples. Moreover, clones with genotypes that are common in more cells are closer to the root of the Conifer inferred tree.

In order to evaluate the performance of Conifer method, comprehensive sets of single-cell and bulk sequencing data are simulated with varying numbers of SNVs, cells, bulk samples, and clones. Additionally, a wide range of error rates, assortment biases, and doublets are considered. By studying simulated datasets, it is shown that Conifer is more accurate than other existing methods on different criteria for evaluation of clone identification and clonal evolutionary tree. For assessing Conifer performance on real datasets, data of a patient with colorectal cancer is used. In this investigation, Conifer provides the genotype of clones, cell population, and clonal tree by considering combined single-cell and bulk sequencing data of the primary and metastatic tumors. In the obtained clonal tree, the evolutionary stage in which the metastasis has occurred is clearly identified.

In conclusion, Conifer provides the clonal tree of tumor heterogeneity by combining single-cell and bulk sequencing data which the former is used for resolving the challenge of identifying similar prevalent clones that co-occur in the tumor and also resolving the ambiguities of phylogeny inference, and the latter is used to reduce the effects of false-negative rate and sampling biases. In addition to resolving these challenges, Conifer provides higher accuracy than other existing methods in clone identification and phylogeny inference.

## Material and method

Conifer provides a Bayesian nonparametric model for inferring clonal trees without any knowledge of clones or their evolutionary tree. In fact, as a prior of this Bayesian nonparametric model, the nested CRP introduced in the study of Blei [33] for defining a "hierarchical topic model" is used with some modifications. In addition, inference of posterior distribution is performed by the MCMC method for approximating distributions over trees, clones, and SNVs allocations. Moreover, the posterior samples of SNVs level allocations for Gibbs sampler are summarized by the Maximum Posterior Expected Adjusted Rand (MPEAR) method [34]. The main components of Confier method are briefly explained below.

**Hierarchical topic model**: The objective of the hierarchical topic model in the study of Blei [34] is identifying subsets of words that co-occur within documents as topics and arranging them into a tree in such a way that more general topics are near to the root. Moreover, a document is a path in the tree, which is generated by the topics that appeared on it.

**Nested CRP:** It is a process for providing a prior on tree topologies without any limitation on its width and depth. To understand nested CRP, the Chinese Restaurant Process (CRP) should be defined first. The CRP is a stochastic process for introducing the distribution of customers, which sequentially enter a restaurant with infinite tables and sit at a randomly selected table. The resulted in sitting plans represent customer clustering. The nested CRP is an extended CRP in which instead of having only one restaurant, it is assumed that there are infinite numbers of Chinese restaurants with infinite numbers of tables. A restaurant is selected as the root in which on each table, there is a card with the name of the restaurant that those who are sitting on that table should go to the next night. In fact, as each restaurant is referred to only once, so they show a tree structure. Therefore, the nested CRP provides a prior on tree topologies and each node of the tree provides a CRP over its descendant.

**Distance dependent CRP:** This process provides a class of distributions over partitions and allows different dependencies between data. In order to consider different dependencies, distance-dependent CRP joins customers to other customers, instead of sitting them at different tables. It implies that if two customers have access to each other through a series of customer connections, then they are sitting at the same table. For representing customer connections, a graph is defined in which nodes and edges represent customers and their connections, respectively. In order words,

if $c_i$ is the index of a customer joining the customer $i$, then the binary $(i, c_i)$ is the directional graph edge. The clusters are defined according to connected components in this similarity graph. additionally, $l(c)$ indicates the label of the cluster for each customer.

**Conifer model description:** Following the hierarchical topic model idea, SNVs, clones, and clonal tree in Conifer method correspond to words, topics, and topic hierarchy, respectively. In addition, a single cell profile corresponds to a document that is generated by the clones on a path in the tree. In fact, each clone, which is a node in the tree, is a probability distribution on the SNVs, and a path is an infinite set of them. In Conifer, the nested CRP model is extended in such a way that instead of ordinary CRP, distance-dependent CRP [35] is used in each node of the tree to define prior over its descendant.

Introducing a clonal tree in Conifer is based on identifying single-cell mutation profiles on the paths generated by the nested CRP. Conifer introduces a two-dimensional generative model that firstly, defines nodes as probability distributions over SNVs and secondly, defines a probability distribution on a set of nodes on each path in the tree. Using notation of Baldassano's study [36] for connectivity clustering model, Conifer's generative model is described as follows:

$$c_d \sim nCRP(\gamma) \qquad \textit{for each column in matrix M (each cell)}\ d \in \{1, \dots, m\} \tag{1}$$

$$z_d \sim ddCRP(\eta, f, t, V_d) \tag{2}$$

$$A_{l_1,l_2}, \sigma^2_{l_1,l_2} \sim Normal - Inverse - \chi^2\ (\mu_0, \kappa_0, {\sigma_0}^2, \nu_0) \tag{3}$$

$$V_{d,ij} \sim Normal(A_{l(z)_i, l(z)_j}, \sigma^2_{l(z)_i, l(z)_j}) \tag{4}$$

$$f(t_{ij}) = \frac{\exp(-t_{ij} + a)}{(1 + exp(-t_{ij} + a))} \tag{5}$$

$$\gamma \sim Gamma(\kappa_\gamma, \theta_\gamma)$$

$$\eta \sim Exponential(\lambda_\eta)$$

$$a \sim Exponential(\lambda_a)$$

Suppose M is a $n \times m$ matrix in which each row and column represents SNV and cell, respectively and its elements with a value of zero show that no mutation has occurred in the corresponding position, while value of one means that the mutation has occurred.

In addition, $B$ is a $n \times b$ matrix that its rows are SNVs and its columns are bulk samples and $B_{ij}$ is a variant allele frequency that corresponds to the i[th] SNV in the j[th] bulk sample. Also, $B_d$ is a submatrix of $B$ which is formed by selecting rows of matrix $B$ corresponding to the mutant loci in cell d. Also, $V_{d,ij}$ is the observed connectivity between $i$ and $j$ loci which is obtained by the Euclidean distance of the i[th] and j[th] rows in $B_d$ submatrix.

In this model, $c_d$ is a tree-structure path generated by nested CRP with a parameter γ following the Gamma distribution for n mutations of the cell d and $z_d$ is the level assignment for cell d which is generated by a distance-dependent CRP. It should be noted that in $z_d$ nodes are ordered according to their mean VAF value. In other words, for each node the higher the mean VAF value, the lower the level number (see Fig. 10).

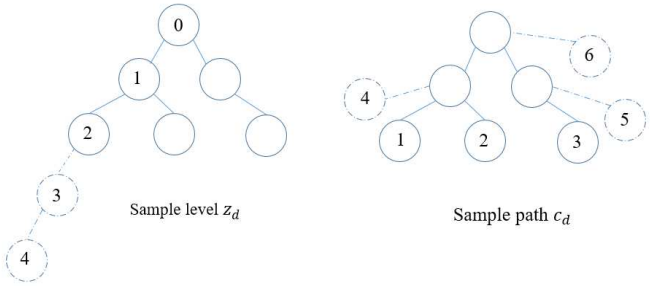



**Fig. 10** Sample level and path for SNVs of cell d

Moreover, $\eta$ is the model parameter following exponential distribution and controls the self-loop in the connectivity graph, and is similar to the probability of selecting a new table in the CRP model. The decay function is represented by $f$ and $t$ is a $n \times n$ matrix and $t_{ij}$ denotes the co-occurrence of two mutations on different cells divided by the total number of cells. In addition, $A$ denotes the connectivity strength with scalar prior mean $\mu_0$ and precision $\kappa_0$ for two $l_1$ $and$ $l_2$ clusters, and $\sigma^2$ is their connectivity variance, with scalar prior mean ${\sigma_0}^2$ and precision $\nu_0$ and determines the size of clusters. The smaller variances result in smaller cluster sizes. $A$ and $\sigma^2$ follow the $Normal - Inverse - \chi^2$ distribution function. The values of the parameters for all experiments are as follows:

$$\eta = 10\,, \mu_0 = 0, \kappa_0 = 0.0001\,, \upsilon_0 = 1$$

Finally, $l(z)$ is the cluster assignment derived from the customer assignment for each locus in each level. For sampling from the posterior distribution, Gibbs sampling is used which is explained in the next section.

**Inference:** Clonal tree of the tumor heterogeneity is found by posterior distribution inference on the path and level assignment of loci from single-cell and bulk sequencing data, which is shown by $p(c_{1:M}, z_{1:M}|\gamma, \eta, f, t)$. The posterior is approximated with Collapsed Gibbs sampling by iterating between sampling level assignments and sampling paths:

1) Sampling level assignments:

$$p(z_{d,n}|z_{-(d,n)}, c, V_d, \eta, f, t) \propto p(z_{d,n}|\eta, f, t)p(V_{d,1:n}|\, l(z_{d,1:n}), c) \tag{6}$$

$$p(z_{d,n}|\eta, f, t) \propto \begin{Bmatrix} f(t_{i,j}) & j \neq i \\ \eta & j = i \end{Bmatrix}$$

$$p(V_{d,1:n}|\, l(z_{d,1:n}), c) \propto \prod_{k_1,k_2}^{unique(z_{d,1:n})} p(V_{d,(z_{d,k_1} z_{d,k_2})})$$

where $unique(z_{d,1:n})$ denotes a unique cluster of level assignment and $v_{d,1:n=k}$ is the loci assigned to the k[th] cluster. The details of $p(V_{d,(z_{d,k_1} z_{d,k_2})})$ calculation as the marginal likelihood of $Normal - Inverse - \chi^2$, are provided in the study of Baldassano [36].

2) Sampling path:

$$p(c_d|c_{-d}, w, z, \gamma, \eta) \propto p(c_d|c_{-d}, \gamma)p(w_d|c, w_{-d}, z, \eta) \tag{7}$$

$$p(c_d|c_{-d}, \gamma) \propto \begin{Bmatrix} |n_i| & j \neq i \\ \gamma & j = i \end{Bmatrix}$$

$$p(w_d|c, w_{-d}, z, \eta) = \prod_{k=1}^{max(z_d)} p(w_{-d}|z_{d,k}, c_{-d}, \eta)$$

$$= \prod_{k=1}^{max(z_d)} \frac{\Gamma(\phi_{z_{d,k},-d}(.) + n)}{\prod_w \Gamma(\phi_{z_{d,k},-d}(w) + \eta)} \frac{\prod_w \Gamma(\phi_{z_{d,k},-d}(w) + \phi_{z_{d,k},d}(w) + \eta)}{\Gamma(\phi_{z_{d,k},-d}(.) + \phi_{z_{d,k},d}(.) + n)}$$

In relation (7), $p(w_d|c, w_{-d}, z, \eta)$ calculates the probability that $w_d$, the column d of matrix M, has created a specific path, and $p(c_d|c_{-d}, \gamma)$ computes its prior. The standard gamma function is shown by $\Gamma$ and $\phi_{z_{d,k},-d}(w)$ denotes the number of SNVs which are assigned to the clone with index $z_{d,k}$ and are not in cell d.

## Abbreviations

**DNA:** Deoxyribonucleic Acid
**CRP:** Chinese Restaurant Process
**CNV:** Copy Number Variation
**CASet:** Common Ancestor Set Distance
**DISC:** Distinctly Inherited Set
**MCMC:** Markov Chain Monte Carlo
**MPEAR**: Maximum Posterior Expected Adjusted Rand
**VAF:** Variant Allele Frequency
**SNV:** Single Nucleotide Variation

## Declarations

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Availability of data and materials

The sequencing datasets analysed during the current study are available from the Sequence Read Archive with the accession number SRP074289.

### Competing interests

The authors declare that they have no competing interests.

### Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

**Authors' contributions**

**LBA** designing and implementing the method, collecting and analyzing the data, and writing-manuscript. **BG and SG** conceptualization, interpreting the result, supervision, project administration, and editing the manuscript. **MFA and SPS**: conceptualization, validation, editing-manuscript. All authors read and approved the final manuscript.

**Acknowledgments**:
None

**Reference**

1. Nowell PC: **The clonal evolution of tumor cell populations**. *Science* 1976, **194**(4260):23-28.
2. Marte B: **Tumour heterogeneity**. In.: Nature Publishing Group; 2013.
3. Marusyk A, Almendro V, Polyak K: **Intra-tumour heterogeneity: a looking glass for cancer?** *Nature Reviews Cancer* 2012, **12**(5):323-334.
4. Merlo LM, Pepper JW, Reid BJ, Maley CC: **Cancer as an evolutionary and ecological process**. *Nature reviews cancer* 2006, **6**(12):924-935.
5. Burrell RA, Swanton C: **Tumour heterogeneity and the evolution of polyclonal drug resistance**. *Molecular oncology* 2014, **8**(6):1095-1111.
6. Greaves M: **Evolutionary determinants of cancer**. *Cancer discovery* 2015, **5**(8):806-820.
7. Dagogo-Jack I, Shaw AT: **Tumour heterogeneity and resistance to cancer therapies**. *Nature reviews Clinical oncology* 2018, **15**(2):81.
8. Kuipers J, Jahn K, Beerenwinkel N: **Advances in understanding tumour evolution through single-cell sequencing**. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer* 2017, **1867**(2):127-138.
9. Roth A, Khattra J, Yap D, Wan A, Laks E, Biele J, Ha G, Aparicio S, Bouchard-Côté A, Shah SP: **PyClone: statistical inference of clonal population structure in cancer**. *Nature methods* 2014, **11**(4):396-398.
10. Popic V, Salari R, Hajirasouliha I, Kashef-Haghighi D, West RB, Batzoglou S: **Fast and scalable inference of multi-sample cancer lineages**. *Genome biology* 2015, **16**(1):1-17.
11. Deshwar AG, Vembu S, Yung CK, Jang GH, Stein L, Morris Q: **PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors**. *Genome biology* 2015, **16**(1):1-20.

12. Jiao W, Vembu S, Deshwar AG, Stein L, Morris Q: **Inferring clonal evolution of tumors from single nucleotide somatic mutations**. *BMC bioinformatics* 2014, **15**(1):1-16.
13. Zare H, Wang J, Hu A, Weber K, Smith J, Nickerson D, Song C, Witten D, Blau CA, Noble WS: **Inferring clonal composition from multiple sections of a breast cancer**. *PLoS Comput Biol* 2014, **10**(7):e1003703.
14. El-Kebir M, Oesper L, Acheson-Field H, Raphael BJ: **Reconstruction of clonal trees and tumor composition from multi-sample sequencing data**. *Bioinformatics* 2015, **31**(12):i62-i70.
15. Griffith M, Miller CA, Griffith OL, Krysiak K, Skidmore ZL, Ramu A, Walker JR, Dang HX, Trani L, Larson DE: **Optimizing cancer genome sequencing and analysis**. *Cell systems* 2015, **1**(3):210-223.
16. Wang Y, Navin NE: **Advances and applications of single-cell sequencing technologies**. *Molecular cell* 2015, **58**(4):598-609.
17. Navin NE: **Cancer genomics: one cell at a time**. *Genome biology* 2014, **15**(8):1-13.
18. Roth A, McPherson A, Laks E, Biele J, Yap D, Wan A, Smith MA, Nielsen CB, McAlpine JN, Aparicio S: **Clonal genotype and population structure inference from single-cell tumor sequencing**. *Nature methods* 2016, **13**(7):573-576.
19. Kuipers J, Jahn K, Raphael BJ, Beerenwinkel N: **Single-cell sequencing data reveal widespread recurrence and loss of mutational hits in the life histories of tumors**. *Genome Research* 2017, **27**(11):1885-1894.
20. Kim KI, Simon R: **Using single-cell sequencing data to model the evolutionary history of a tumor**. *BMC bioinformatics* 2014, **15**(1):1-13.
21. Malikic S, Jahn K, Kuipers J, Sahinalp SC, Beerenwinkel N: **Integrative inference of subclonal tumour evolution from single-cell and bulk sequencing data**. *Nature communications* 2019, **10**(1):1-12.
22. Jahn K, Kuipers J, Beerenwinkel N: **Tree inference for single-cell data**. *Genome biology* 2016, **17**(1):1-17.
23. Ross EM, Markowetz F: **OncoNEM: inferring tumor evolution from single-cell sequencing data**. *Genome biology* 2016, **17**(1):1-14.
24. Zafar H, Tzen A, Navin N, Chen K, Nakhleh L: **SiFit: inferring tumor trees from single-cell sequencing data under finite-sites models**. *Genome biology* 2017, **18**(1):1-20.
25. Zafar H, Navin N, Chen K, Nakhleh L: **SiCloneFit: Bayesian inference of population structure, genotype, and phylogeny of tumor clones from single-cell genome sequencing data**. *Genome Research* 2019, **29**(11):1847-1859.
26. Salehi S, Steif A, Roth A, Aparicio S, Bouchard-Côté A, Shah SP: **ddClone: joint statistical inference of clonal populations from single-cell and bulk tumour sequencing data**. *Genome biology* 2017, **18**(1):1-18.
27. Rosenberg A, Hirschberg J: **V-measure: A conditional entropy-based external cluster evaluation measure**. In: *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL): 2007*. 410-420.
28. Hubert L, Arabie P: **Comparing partitions**. *Journal of classification* 1985, **2**(1):193-218.
29. DiNardo Z, Tomlinson K, Ritz A, Oesper L: **Distance measures for tumor evolutionary trees**. *Bioinformatics* 2020, **36**(7):2090-2097.
30. Hadley W: **Ggplot2: Elegrant graphics for data analysis**: Springer; 2016.

31. Leung ML, Davis A, Gao R, Casasent A, Wang Y, Sei E, Vilar E, Maru D, Kopetz S, Navin NE: **Single-cell DNA sequencing reveals a late-dissemination model in metastatic colorectal cancer**. *Genome Research* 2017, **27**(8):1287-1299.
32. Satas G, Zaccaria S, Mon G, Raphael BJ: **Scarlet: Single-cell tumor phylogeny inference with copy-number constrained mutation losses**. *Cell Systems* 2020, **10**(4):323-332. e328.
33. Blei DM, Griffiths TL, Jordan MI: **The nested Chinese restaurant process and bayesian nonparametric inference of topic hierarchies**. *Journal of the ACM (JACM)* 2010, **57**(2):1-30.
34. Fritsch A, Ickstadt K: **Improved criteria for clustering based on the posterior similarity matrix**. *Bayesian analysis* 2009, **4**(2):367-391.
35. Blei DM, Frazier PI: **Distance dependent Chinese restaurant processes**. *Journal of Machine Learning Research* 2011, **12**(8).
36. Baldassano C, Beck DM, Fei-Fei L: **Parcellating connectivity in spatial maps**. *PeerJ* 2015, **3**:e784.

# Figures

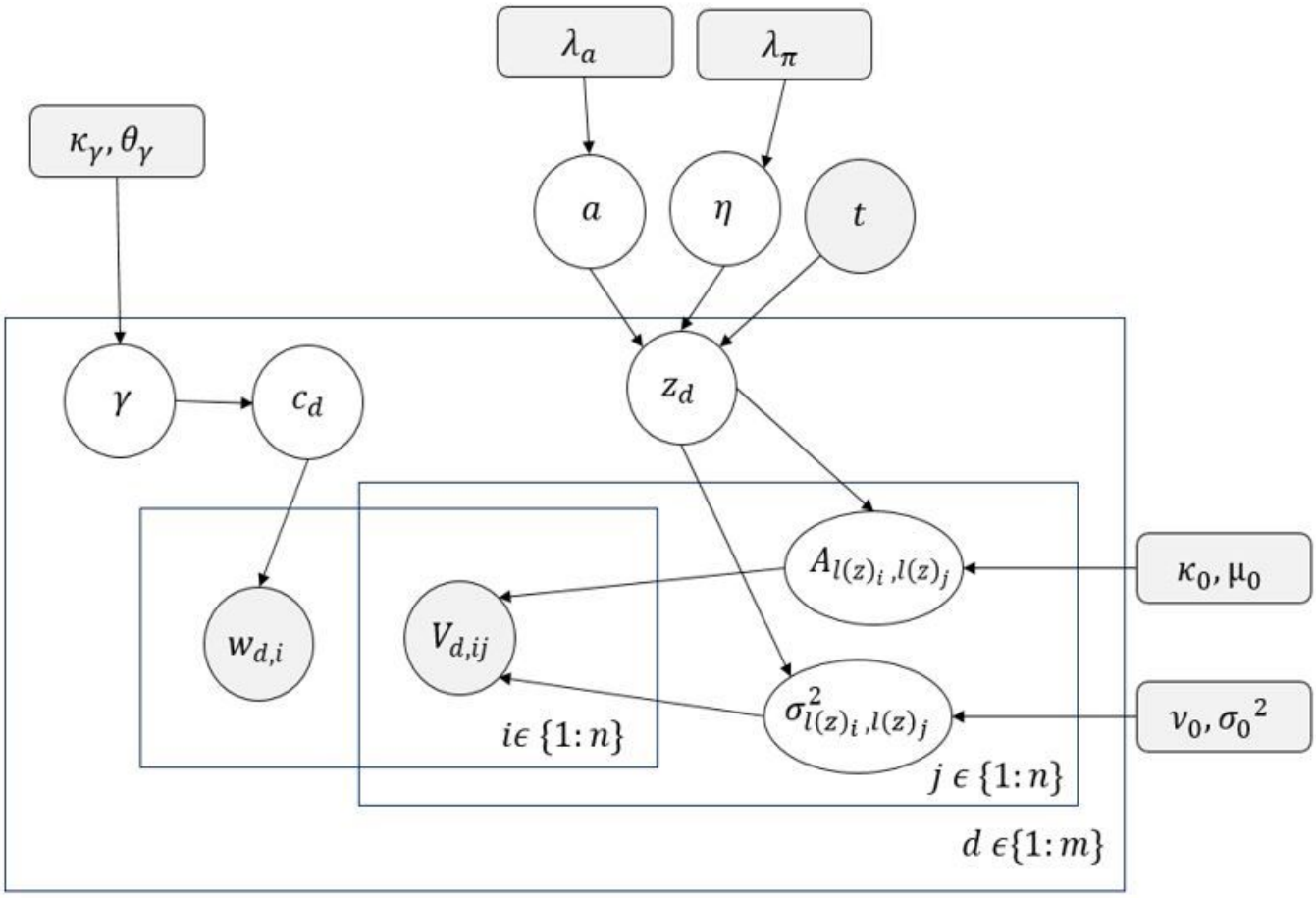
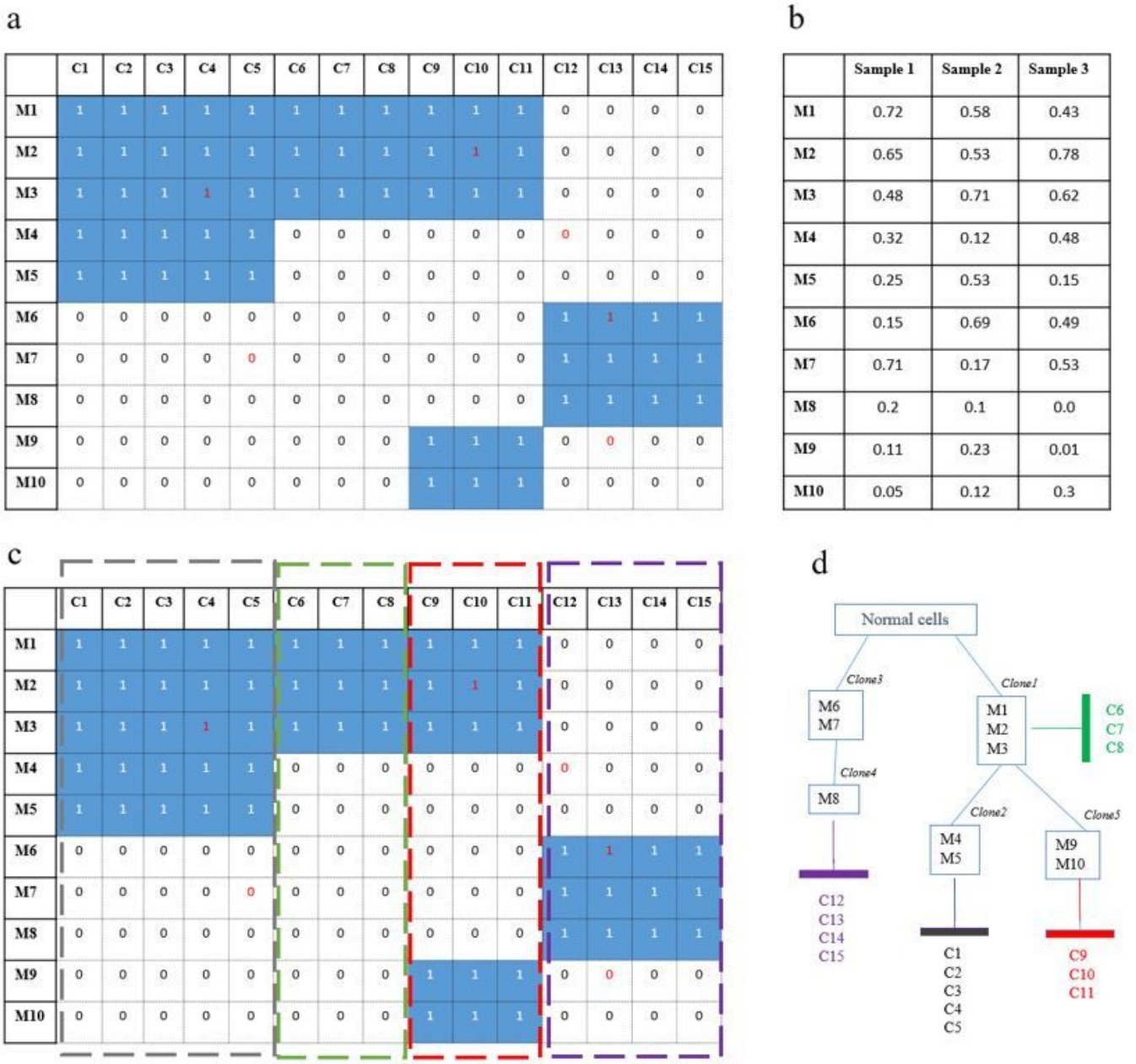



**Figure 1**

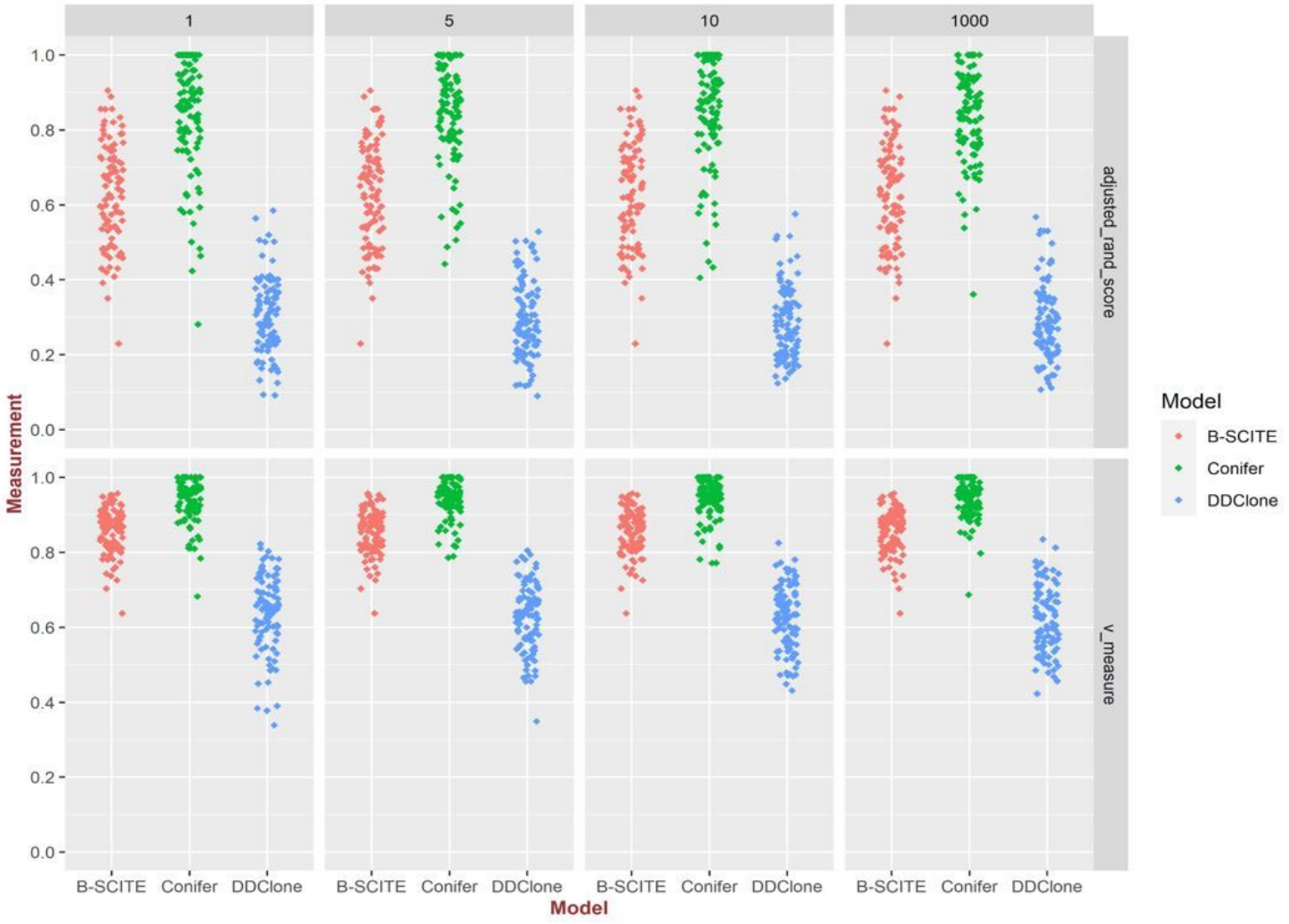
please see the manuscript file for the full caption




| | Sample 1 | Sample 2 | Sample 3 |
|---|---|---|---|
| M1 | 0.72 | 0.58 | 0.43 |
| M2 | 0.65 | 0.53 | 0.78 |
| M3 | 0.48 | 0.71 | 0.62 |
| M4 | 0.32 | 0.12 | 0.48 |
| M5 | 0.25 | 0.53 | 0.15 |
| M6 | 0.15 | 0.69 | 0.49 |
| M7 | 0.71 | 0.17 | 0.53 |
| M8 | 0.2 | 0.1 | 0.0 |
| M9 | 0.11 | 0.23 | 0.01 |
| M10 | 0.05 | 0.12 | 0.3 |

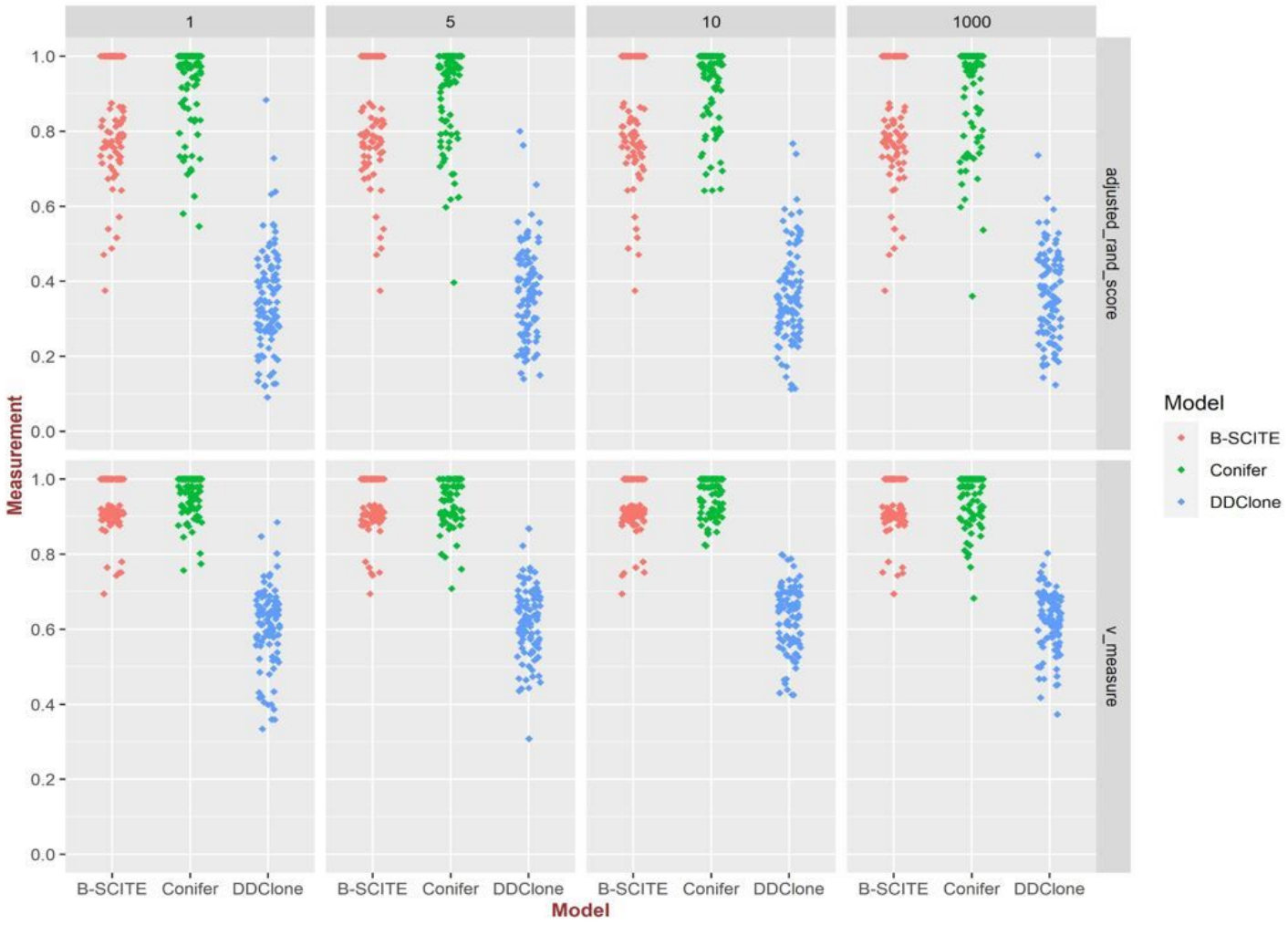
## Figure 2

Schematic representation of combining single-cell and bulk sequencing data for clonal tree inference in Conifer method. a) n×m matrix in which each row and column represents SNVs and cell, respectively. White elements show no mutation and blue ones show mutation has occurrence. 1 and 0 with the red font show false-positive and false negative (drop-out events), respectively. b) n×b matrix that its rows are SNVs and its columns are bulk samples and B_ij is variant allele frequency in bulk samples. c) co-occurred patterns of SNVs in single-cell profiles which are determined by dashed rectangles. d) the clonal tree and cell population attachment.



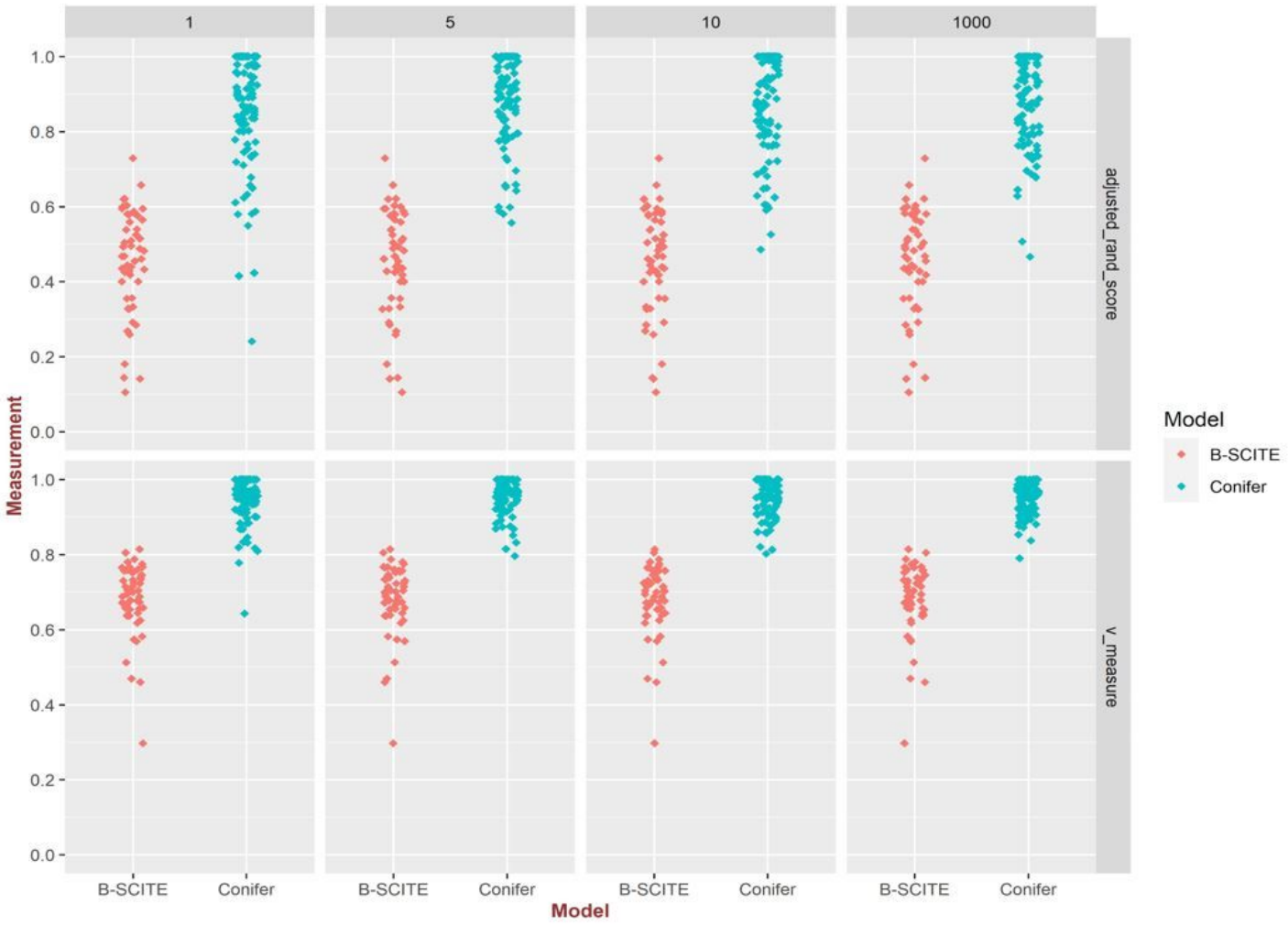

## Figure 3

Comparison of mutation clustering accuracy in ddClone, B-SCITE, and Conifer methods for 100 clonal trees simulated with 10 clones and 50 mutations. For λ = 1,5,10 and 1000. For single-cell data, 50 genotypes are extracted for each clonal tree. The number of bulk sequencing samples is 1 with coverage 10,000. The following errors are added to the single-cell set: the false-positive rate of 10^(-5), the false-negative rate of 0.2, missing rate of 0.05, and doublet rate of 0.2.



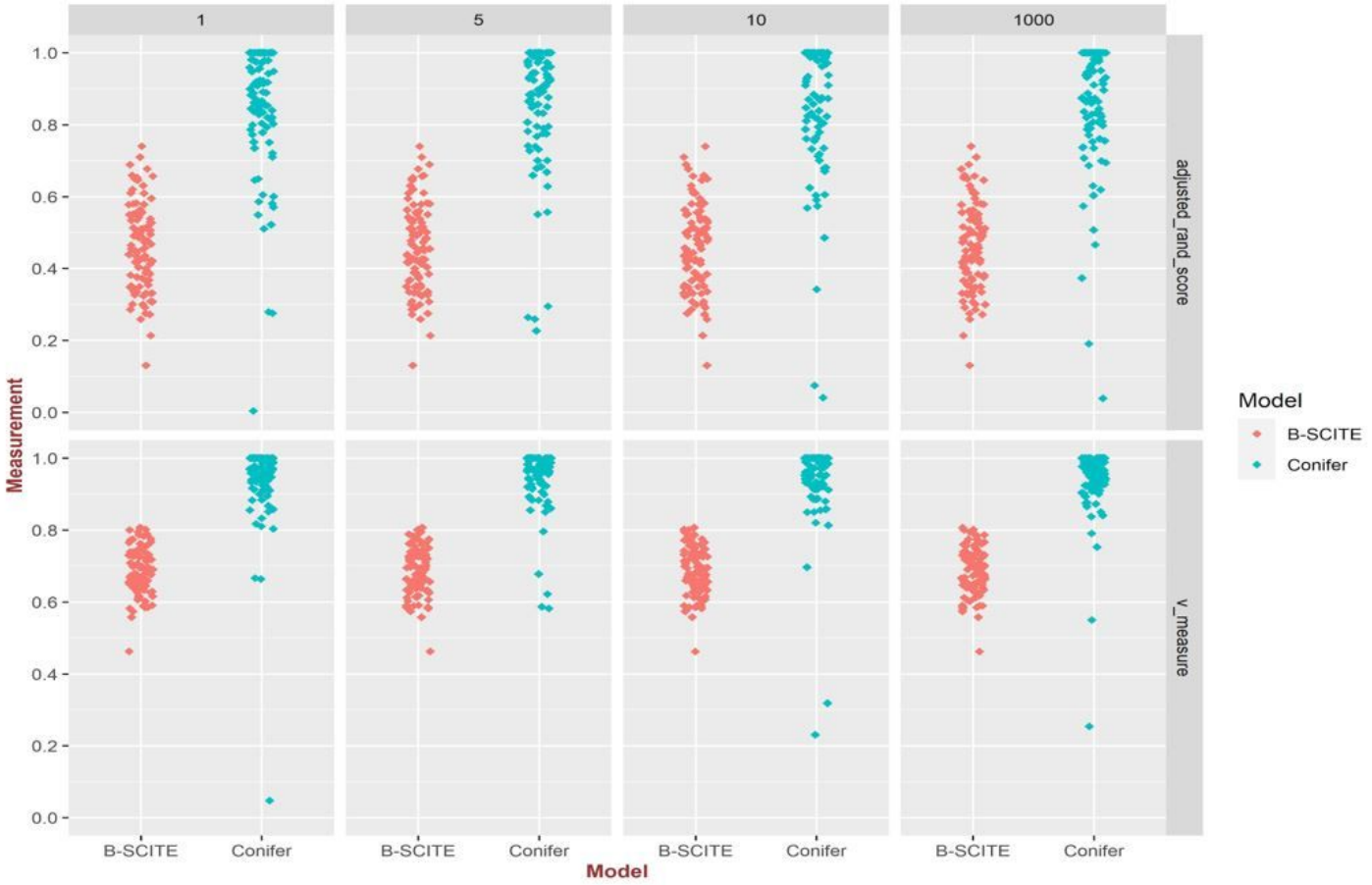

## Figure 4

Comparison of mutation clustering accuracy in ddClone, B-SCITE, and Conifer methods for 100 clonal trees simulated with 6 clones and 50 mutations. For λ = 1,5,10 and 1000. For single-cell data, 50 genotypes are extracted for each clonal tree. The number of bulk sequencing samples is 1 with coverage 10,000. The following errors are added to the single-cell set: the false-positive rate of 10^(-5), the false-negative rate of 0.2, missing rate of 0.05, and doublet rate of 0.2.



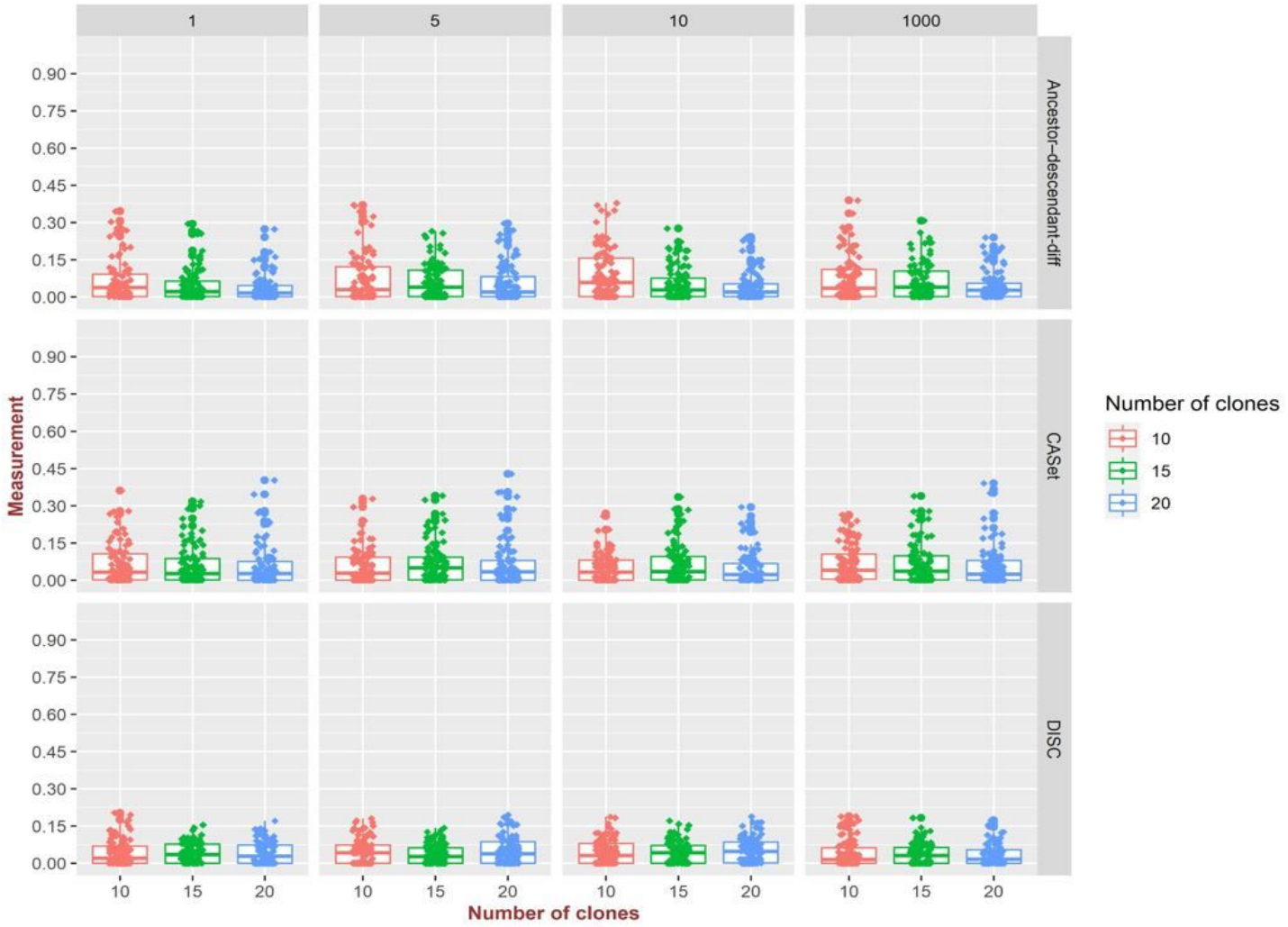

## Figure 5

Comparison of mutation clustering accuracy in B-SCITE and Conifer methods for 100 clonal trees simulated with 10 clones and 50 mutations. For λ equals 1,5,10 and 1000. For single-cell data, 50 genotypes are extracted for each clonal. The number of bulk sequencing samples is 2 with coverage 10,000. The following errors are added to the single-cell set: the false-positive rate of 10^(-5), the false-negative rate of 0.2, missing rate of 0.05, and doublet rate of 0.2.



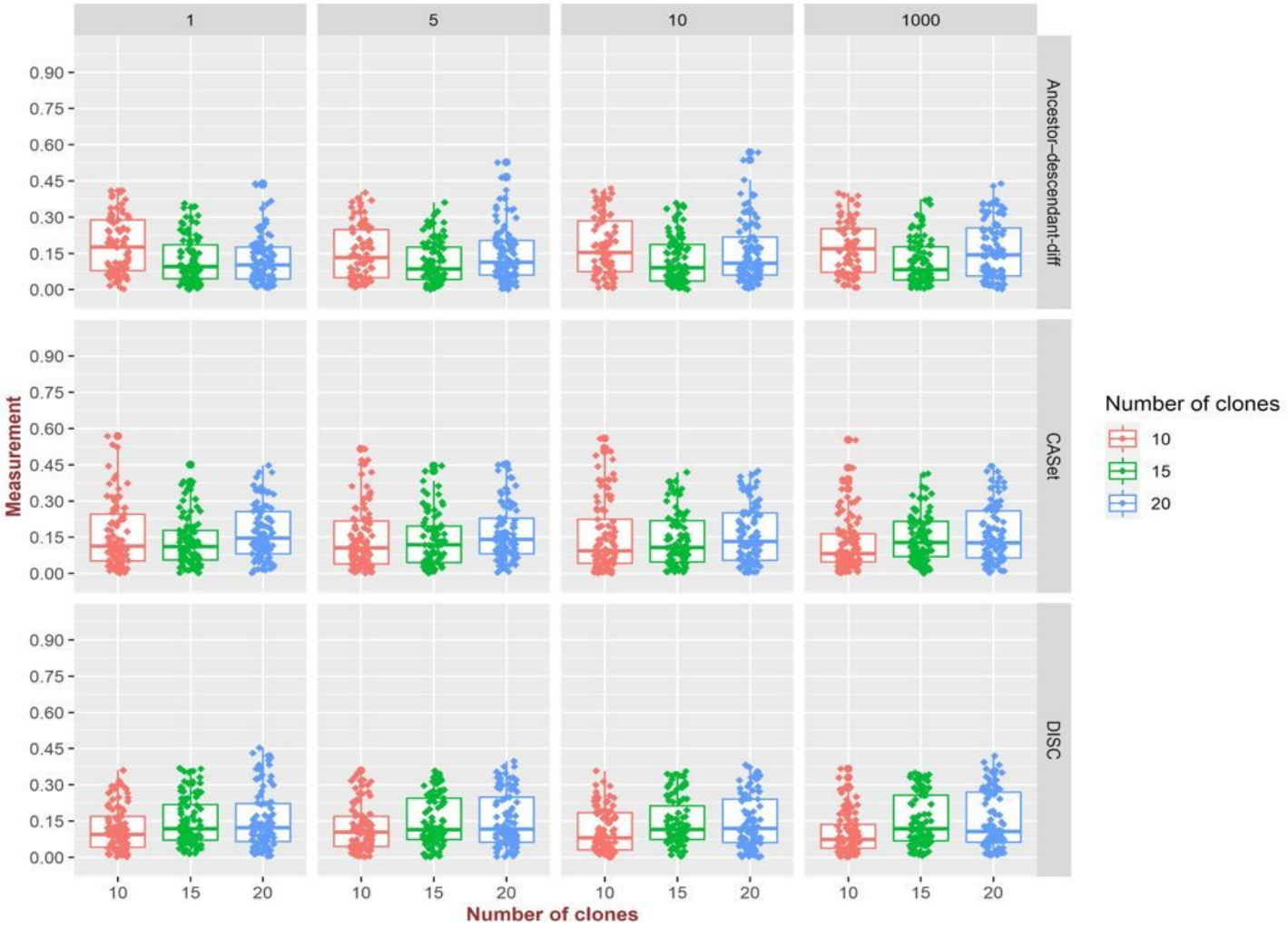

## Figure 6

Comparison of mutation clustering accuracy in B-SCITE and Conifer methods for 100 clonal trees simulated with 10 clones and 50 mutations. For λ equals to 1,5,10 and 1000. For single-cell data, 50 genotypes are extracted for each clonal tree. The number of bulk sequencing samples is 3 with coverage 10,000. The following errors are added to the single-cell set: the false-positive rate of 10^(-5), the false-negative rate of 0.2, missing rate of 0.05, and doublet rate of 0.2.




## Figure 7

The clonal tree distances of Conifer for 100 clonal trees simulated with 10,15 and 20 clones with 100 mutations For λ = 1,5,10 and 1000. For single-cell data, 100 genotypes are extracted for each clonal tree. The number of bulk sequencing samples is 3 with coverage 10,000. The following errors are added to the single-cell set: the false-positive rate of 10^(-5), the false-negative rate of 0.2, missing rate of 0.03, and doublet rate of 0.2.




## Figure 8

The clonal tree distances of Conifer for 100 clonal trees simulated with 10,15 and 20 clones with 100 SNVs for λ = 1,5,10 and 1000. 15 percent of SNVs are selected from copy number change regions. For single-cell data, 100 genotypes are extracted for each clonal tree. The number of bulk sequencing samples is 3 with coverage 10,000. The following errors are added to the single-cell set: the false-positive rate of 10^(-5), the false-negative rate of 0.2, missing rate of 0.03, and doublet rate of 0.2.

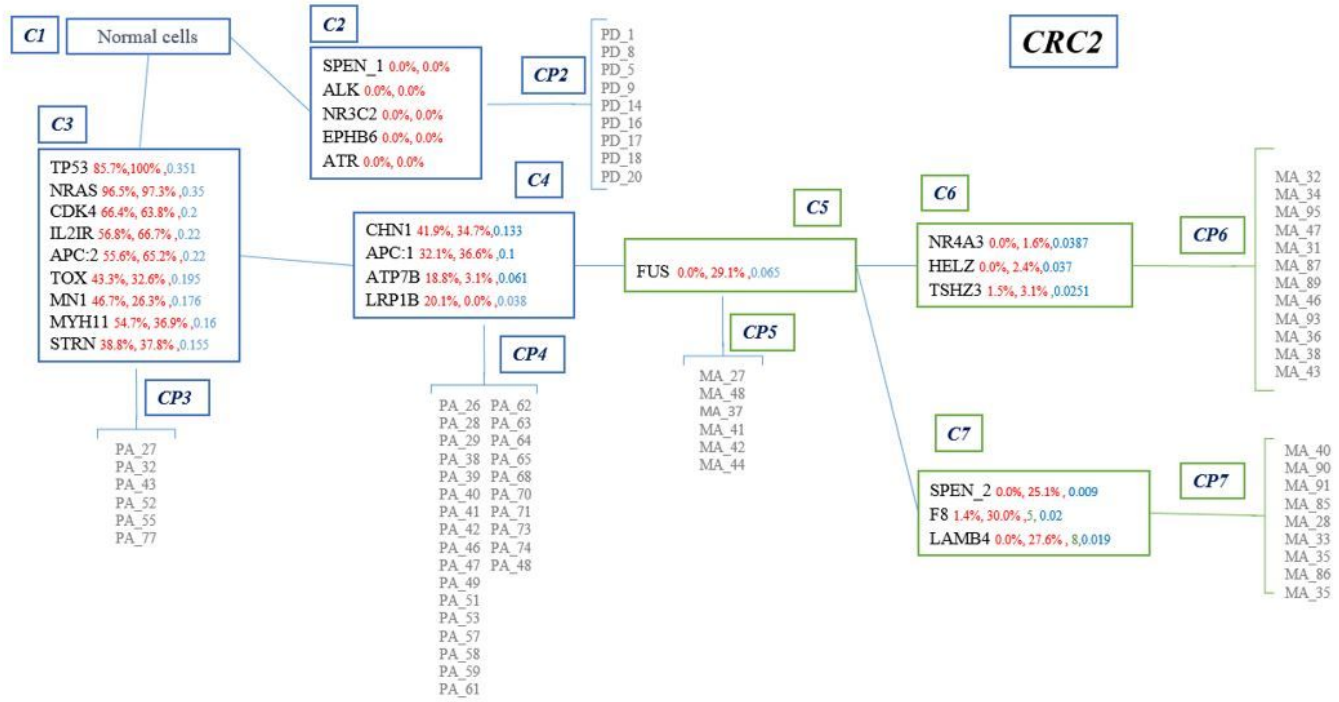



**Figure 9**

Clonal evolution tree for CRC2 patient tumor data. For each SNV, three numbers are reported, from left, the first and second numbers are the VAFs in colorectal tumor bulk sample and metastasis liver bulk sample, respectively, and the third number is the frequency of that SNV in the single-cell sequencing data

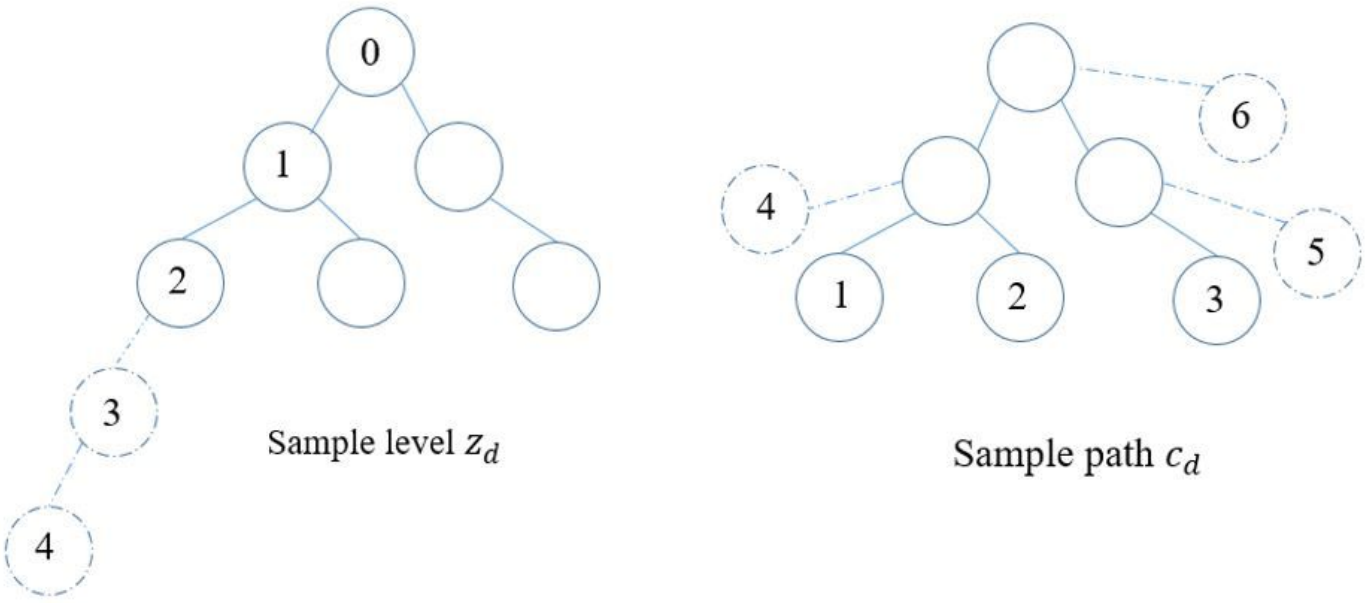



**Figure 10**

Sample level and path for SNVs of cell d